DOCUMENT RESUME

ED 409 342 TM 026 825

AUTHOR Jiang, Ying Hong; And Others
TITLE Error Sources Influencing Performance Assessment Reliability or Generalizability: A Meta Analysis.
PUB DATE Mar 97
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Error of Measurement; Estimation (Mathematics); *Generalizability Theory; Judges; Meta Analysis; *Performance Based Assessment; Research Reports; *Test Reliability
IDENTIFIERS *Experts

ABSTRACT

As performance-based assessments have gained wider use, there are increasing concerns about their dependability. This study is a synthesis of existing studies regarding the reliability or generalizability of performance assessments. The meta-analysis involves summarizing, examining, and evaluating research findings. Articles on the dependability of performance assessments, analyzed through traditional means or a generalizability framework published after 1980 were selected. The literature search yielded 22 studies meeting the criteria for inclusion. These 22 studies yielded 258 different reliability or generalizability coefficients. Task and occasion facets contributed the greatest proportion of variance to estimates of error in the measurement procedure. Both are inherent in the construction of many performance tasks. The judge facet did not contribute a large proportion of error variance. Critics of performance assessment should not worry that the use of professional judgment to score performance assessment will be a major source of measurement error. (Contains 3 tables and 25 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 409 342

Error Sources Influencing Performance Assessment Reliability or Generalizability: A Meta Analysis

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Ying Hong Jiang
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Ying Hong Jiang

Philip L. Smith

Paul Nichols

University of Wisconsin-Milwaukee

TM026825

BEST COPY AVAILABLE

Error Sources Influencing Performance Assessment Reliability or Generalizability: A Meta Analysis

I. Introduction:

Today, there is wide recognition of the important roles performance-based assessments are playing in the field of measurement and assessment. However, as performance-based assessments have gained wider use, concerns have arisen regarding their dependability. Studies investigating the reliability or generalizability of performance assessments have yielded inconsistent and often discouraging results. While some studies have shown that particular performance assessments have reached acceptable levels reliability (Klein and Seligsohn, 1987), more commonly studies show low reliability of performance assessments. These findings raise concerns that performance assessments will not meet acceptable technical standards for providing dependable measurement (Koretz, McCafferey, Klein, Bell, & Stecher, 1992; Shavelson, Gao, & Baxter, 1993).

Conceptually, a performance assessment score may be viewed as a sample of student performance drawn from a complex universe defined by a combination of all admissible tasks, occasions, raters, and measurement methods. Generalizability theory refers to each of these dimensions as "facets". The task facet represents the content in a subject-matter domain; The occasion facet includes all possible occasions on which a decision maker would be equally willing to accept a score on the performance assessment; The rater facet includes all possible individuals who are trained to score the performance reliably. Typically, these

three facets are viewed as primary sources of error in a measurement procedure, especially a performance assessment.

However, researchers frequently find inconsistent, sometimes even contradictory results regarding which of these sources most influences the reliability or generalizability of performance assessment. While some studies (Koretz, 1992) have estimated approximately 10 percent of the total variance attributable to raters, other studies (Shavelson, Gao, and Baxter, 1993) have indicated that the estimated variance accounted for by raters was near zero. Other studies (Shavelson, Gao, and Baxter, 1993) have shown that task sampling contributes the most error to performance assessments.

Given the importance placed on the reliability of a measurement procedure and the mixed findings regarding the reliability of performance assessments, it seems compelling and appropriate to conduct a synthesis on the existing studies regarding the reliability or generalizability for performance assessments, even though that literature is somewhat limited. The purpose of this study is to investigate and identify the sources of poor reliability for performance assessment through reviewing and synthesizing existing studies of reliability or generalizability. This meta-analysis involves summarizing, examining, and evaluating or re-analyzing research findings to reach some general conclusions regarding effects of a given source. In addition, we will examine closely studies which present exceptions to these conclusions in an attempt to

understand the contexts which lead to more or less reliable performance-based assessment.

II. Method:

Commonly used procedures for meta-analytic research (Glass, McGaw, and Smith, 1981) require researchers to: (a) locate studies of an issue through objective and replicable searches, (b) code the studies for salient features, (c) describe study outcomes on a common scale, and (d) use summary methods (possible statistical) to find relationships between study features and study outcomes. In this section, we describe how we carried out each step in our quantitative synthesis of the reliability of performance assessments.

Locating Studies

The following criteria were used to select articles: (1) the study had to employ a performance assessment as the instrument; (2) the study had to study the dependability of the performance assessment instrument through traditional means or a generalizability framework; (3) the study had to report statistical information relating to reliability or generalizability coefficients; (4) the study had to be published in 1980 or later. The sources for the literature review were Educational Resources Information Center (ERIC) and Psychology Literature (PSYCHLIT). Discriptors used to locate studies were 'alternative assessment and reliability', 'authentic assessment and reliability', 'performance assessment and reliability', 'alternative assessment and generalizability', 'authentic

assessment and generalizability', 'performance assessment and generalizability'.

Since the purpose of this study is to investigate and identify the sources related to lower reliability in performance assessments, the study features are coded according to the type of sources of errors related to the study designs. The conceivable sources of errors related to the performance assessment reliability or generalizability include: (a) momentary inattention from the test takers or persons, (b) the particular set of items, (c) the particular set of judges, and (d) the particular occasions. These sources of errors might contribute jointly or separately to errors in the measurement procedure.

A) If a study employs a generalizability analysis, and if the study reports variance components estimates related to each facet involved in the analysis, then a generalizability coefficient is estimated accordingly (Shavelson and Webb, 1991). B) All the generalizability and reliability coefficients were "stepped down" to reflect the situations where one task, one rater or one occasion is involved.

The coefficients reported by the studies selected were coded according to categories of reliability or generalizability coefficients. For the reliability coefficients, categories included internal consistency, interrater, and test-retest. Similarly, the generalizability coefficients were categorized as pxixj (person by items by judges), and pxixjxo (person by items by judges by occasions). Conceivably, there could be more

categories, but these five categories described all the coefficients that included in the current study.

It is important to note at this point that many of the coefficients used in this study are not independent in that multiple coefficients were reported in many studies. These multiple coefficients were often derived from the same subject sample, items or judges and therefore are not statistically independent. Consequently, the summaries that will be drawn from these data are influenced more by those studies reporting higher numbers of these coefficients.

For the coefficients that have been found to be exceptionally high or low, we further examined the contexts of studies in which those coefficients are reported.

Analysis

The analysis was organized around the four sources of error: (a) momentary inattention from the test takers or the persons, (b) particular set of items, (c) set of judges, (d) particular occasions. These sources contribute separately or jointly to the reliability or generalizability coefficients. Means and medians of reliability coefficients or generalizability coefficients were computed for each of the four error sources.

III. Results and Discussion

The search yielded twenty two studies meeting the criteria for inclusion in the current study. It should be noted that the criteria for inclusion in the study and the literature search descriptors eliminated from inclusion in this study the reasonably large body of literature related to writing and

composition, which may, to some, be viewed as "performance assessment". The exclusion of this literature limits the conclusions of this study to content areas other than writing. Specifically, the content areas represented in these studies included language art skills, geometry proof and problem solving skills in mathematics, reasoning skills, teacher performance, vocational education, managerial writing, etc.

There are three studies that reported internal consistency reliability coefficients (Bachman & et al, 1993; Greenan & McCabe, 1989; Klein & Seligsohn, 1987), eleven studies that reported inter-rater reliability coefficients (Barrett, 1992; Burger & Burger, 1994; Cronin & Capie, 1986; Gearhart, et al, 1992; Greenan, 1989; Koretz, et al, 1992; Rogers, 1994; Thompson & Daily, 1991; Trent & Gilman, 1984; Webb, Raymond & Houston, 1990; Zollie, et al, 1990), and one study that reported test-retest reliability (Greenan, 1989). There are eight studies that reported generalizability coefficients with pxixj designs (Bachman, et al, 1993; Capie & Cronin, 1986; Cronin & Capie, 1985; Koretz, et al, 1992; Linn, 1993; Shavelson, Gao & Baxer, 1993; Telese & Kulm, 1995; Yap & Capie, 1985), and three studies that reported generalizability coefficients with pxixjxo designs (Cronin & Capie, 1986; Shavelson, Gao & Baxer, 1993; Tobin & Capie, 1981).

In the sample there are six generalizability studies that provided complete descriptions of the variance component estimates for fully crossed pxixj design and two generalizability

studies that provided complete descriptions of the variance component estimates for fully crossed pxixjxo design.

The twenty two studies have yielded 258 different reliability or generalizability coefficients.

As shown in Table 1, the means are .901, .702, and .801 and medians are .956, .806 and .801 for internal consistency, interrater and test retest reliability coefficients, respectively. The means are .396 and .133, and medians are .314 and .133 for pxixj design and pxixjxo design generalizability coefficients, respectively.

The degree of reduction in the reliability of assessment differs across sources of error. When the interrater reliability coefficient median is compared with the internal consistency reliability median, there is a fifteen percent reduction due to the additional error source from raters. When the test-retest reliability median is compared with the interrater reliability median, there is a slight decrease. However, we should view this decrease with caution, since there is only one case of test-retest reliability. Interestingly, when we compare the pxixj generalizability coefficient median with the interrater reliability coefficient median, there is more than a forty percent reduction due to the additional source from item or task variance. Then when the pxixjxo generalizability coefficient median is compared to the pxixj generalizability coefficient median, there is almost a twenty percent reduction due to the additional source from occasion.

There is a predictable relationship between the number of sources of error accounted for by the design and the magnitude of means or medians of the reliability coefficients or generalizability coefficients. As the number of error sources increases, the magnitude of means or medians of the reliability coefficients decreases.

The median proportions were calculated for the variance component estimates for the six studies which employed fully crossed pxixj design and provided variance components estimates (See Table 2). As Table 2 shows, the two major sources of variance are the variance due to the person facet which is twenty one percent and the variance due to different items or tasks, which is thirteen percent. The person by item interaction turns out to attribute to twenty three percent of the total variance, which is almost equivalent to the amount of variance due to person facet. The two minor sources of variance are the person by judge interaction, which attributes to five percent of the total variance, and the interaction between item and judge, which is only one percent of the total variance. There is almost no variance due to judges or raters. The three-way interaction between person, item and judge plus error contributes the most variance, which is equal to thirty-five percent of the total variance. When we compare the person, item and judge three way interaction with the person and item interaction, twelve percent of the total variance due to judge is inseparably from error.

Similarly, the median proportions were calculated for the variance component estimates for the two studies which employed

fully crossed pxixjxo design and provided variance component estimates (See Table 3). As Table 3 shows for the pxixjxo, there is little variance from the main facets for person, item, judge, or occasion. A major source of variance is the interaction effect for person and item, which is 18 percent of the total variance. Another major source of variance is the person, item and occasion three way interaction effect, which is thirty seven percent of the total variance. When we compare the person, item, judge and occasion plus error variance to person, item, and judge variance, there is twenty-nine percent of the total variance due to the additional source from occasion that is inseparable from error.

We further examined studies which showed little or no task variation. One common characteristic of these studies is that the tasks had been decomposed into smaller more homogeneous tasks. For example, the study by Bachman, Lyle, et al, (1993) investigating the reliability of the Language Ability Assessment System (LAAS) reported that the variance component from the Grammar Scale task was only .003. The whole assessment system consisted of several different scales for pronunciation, vocabulary, cohesion, organization, and grammar. Another study (Capie & Cronin, 1986) that reported little task variation was conducted to investigate the generalizability of Teacher Performance Assessment Instrument (TPAI), which included eight different competencies. It is noticeable that studies reported higher task variations employed designs to investigate dependability of more complex tasks (Linn, 1993, Shavelson, Gao &

Baxter, 1993). These tasks are the kind of tasks called for by proponents of alternative assessment.

IV. Conclusion

Readers familiar with performance assessment are probably not surprised that task and occasion facets contribute the greatest proportion of variance to estimates of error in the measurement procedure. Variance due to differences in task difficulty is inherent in the construction of many performance tasks. Performance tasks often require the test taker to integrate different content and skills. A typical task may require content knowledge, reasoning, critical thinking, and communication skills. The task allows multiple correct solutions so that different test takers may take different paths in responding to the task. Such complex tasks are unlikely to be equally difficult.

Similarly, variance due to occasion is inherent in the construction of many performance tasks. These tasks provide students with a greater opportunity to learn, and more connections with instruction, than multiple-choice items. Partly, this is because performance tasks simply require considerably more time to answer than multiple choice questions. Perhaps more importantly, performance tasks are similar to and sometimes used as instructional activities.

The judge facet did not contribute a large proportion of error variance. Variance due to human judgment can be minimized through rigorous training procedures. Critics of performance assessment should set aside worries that the use of professional

judgment to score performance assessment will be a major source of measurement error.

References

References marked with an asterisk indicate studies included in the meta-analysis.

*Bachman, L. F., Lynch, B. K., & Mason, M., (1993). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking (Report No. FL 021 399). (ERIC Document Reproduction Service No. ED 368 154).

*Burger, S. E., & Burger, D. L. (1994). Determining the validity of performance-based assessment. Educational Measurement: Issues and Practice, Spring, 9-15.

*Capie, W., & Cronin, L. (1986). How many teacher performance criteria should there be (Report No. TM 860 343)? (ERIC Document Reproduction Service No. ED 270 465).

Cronbach L. J.(1984) Essentials of psychological testing. 4th edition. New York, N. Y.: Harper & Row, c1984.

*Cronin, L. L., & Capie, W. (1985). The influence of scoring procedures on assessment decisions and their reliability (Report No. TM 850 502). (ERIC Document Reproduction Service No. ED 265 167).

*Cronin, L., & Capie, W. (1986). The influence of daily variation in teacher performance on the reliability and validity of assessment data (Report No. TM 860 581). (ERIC Document Reproduction Service No. ED 274 704).

Gitomer, D. H. (1995). Performance assessment: The significance of reliability. In R. L. Brennan (Chair), Toward a broader conception of reliability. Symposium conducted at the annual meeting of the American Educational Research Association,

San Francisco.

Glass, G., McGaw, B., & Smith, M. L. (1981). Meta Analysis in Social Research. Beverly Hills, California: Sage Publications.

*Greenman, J. P., & McCabe, C. (1989). Development and validation of generalizable reasoning skills assessment strategies and procedures. Journal of Industrial Teacher Education, 26(3), 38-50.

*James, A, T., & Gerald, K. (1995). Performance-based assessment of at-risk students in mathematics: the effects of context and setting (Report No. TM 023 226). (ERIC Document Reproduction Service No. ED 382 685).

*Klein, R. S., & Seligsohn, H. C. (1987). Reliability of the Florida vocational achievement performance tests. Journal of Industrial Teacher Education, 24(3), 91-93.

*Koretz, D., Stecher, B, Klein, S., McCaffrey, D., & Deibert, E. (1992). Can portfolios assess student performance and influence instruction? The 1991-1992 Vermont experience (Report No. TM 020 884). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing; Santa Monica, CA: Rand Corp. Inst. for Education and Training. (ERIC Document Reproduction Service No. ED 365 699).

*Koretz, D., McCaffrey, D., Klein, S., Bell, R., & Stecher, B. (1992). The reliability of scores from the 1992 Vermont portfolio assessment program (Report No. TM 019 641). (ERIC Document Reproduction Service No. ED 355 284).

*Linn, R. L. (1994). Generalizability of new standards project, 1993 pilot study tasks in mathematics. Technical Issues

in procedures to link state results to a common national standard. Project 2.4 quantitative models to monitor the status and progress of learning and performance and their antecedents. Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing. (ERIC Document Reproduction Service No. ED 379 356).

*Maryl, G., Herman, J., Baker, E., & Whittaker, A.K. (1992). Writing portfolios at the elementary level: a study of methods for writing assessment (Report No. TM 018 201). (ERIC Document Reproduction Service No. ED 344 900).

*Rogers, P. S., (1994). Analytic measures for evaluating managerial writing. Journal of Business and Technical Communication, 8(4), 380-407.

*Shavelson, R. J., Gao, X., & Baxter, G. P. (1993). Sampling variability of performance assessments. Journal of Educational Measurements, 30(3), 215-232.

Shavelson, R. J., Webb, N. M. (1991) Generalizability Theory: A primer. Newbury Park, London, New Delhi: The International Professional Publishers; Beverly Hills, California: Sage Publications.

*Thomas, B. J. (1992). Implementation of an integrated language arts performance assessment in a large urban school district: technical issues in aggregating and reporting results (Report No. TM 019 254). (ERIC Document Reproduction Service No. ED 352 371).

*Thompson, D. W., Daily, D. L., & Ruma, P. R. (1991). Holistic assessment of a high school writing skills curriculum.

(Report No. CS213 166). (ERIC Document Reproduction Service No. ED 341 992).

*Tobin, K. G. Capie, W.(1981). An empirical investigation of the stability of variance components and dependability coefficients derived from teacher performance data (Report No. TM 810 548). (ERIC Document Reproduction Service No. ED 206 656).

*Trent, J. H., & Gilman, R. A.(1984). An adaptation of the teacher performance assessment instruments (TPAI) in a teacher preparation program in northern Nevada (Report No. TM 840 796). (ERIC Document Reproduction Service No. ED 254 532).

*Webb, L., Raymond, M. R., & Houston, W. M. (1990). Rater stringency and consistency in performance assessment. (ERIC Document Reproduction Service No. ED 318 776).

*Yap K. C., & Capie, W. (1985). The influence of same day or separate day observations on the reliability of assessment data (Report No. TM 850 501). (ERIC Document Reproduction Service No. ED 265 166).

*Zollie, S. Jr., & Averrett, C. P., & Vichers, D. (1990). The reliability of using a focused-holistic scoring approach to measure student performance on a geometry proof (Report No. TM 014 832). (ERIC Document Reproduction Service No. ED 319 748).

Table 1

Type of Mean and Median Reliability Generalizability Coefficients Characterized by Conceivable Error Sources

(Articles=22, Number of Coefficients=258)

Source of errors	Internal consistency	Inter-rater	Test-retest	pxixj	pxixjxo
Momentary inattention	X	X	X	X	X
Set of items				X	X
Set of judges		X	X	X	X
Set of occasions			X		X
Mean reliability and / or generalizability coefficients	.901 (n=21)	.702 (n=151)	.801 (n=1)	.396 (n=77)	.141 (n=8)
Median reliability and / or generalizability coefficients	.956 (n=21)	.806 (n=151)	.801 (n=1)	.314 (n=77)	.133 (n=8)

Table 2

Median and Mean Proportion Of Variance Components Estimates

(Design=pxixj)

(Number of articles=6, Number of designs=30)

Source	Median Proportion	Mean Proportion	Range
Person	.206	.208	.905
Item	.131	.145	.431
Judge	.000	.003	.014
Person X Item	.230	.232	.533
Person X Judge	.050	.057	.234
Item X Judge	.006	.009	.038
Person X Item X Judge, Error	.253	.347	.934

Table 3

Median and Mean Proportion Of Variance Components Estimates

(Design=pxixjxo)

(Number of articles=2, Number of designs=2)

Source	Median Proportion	Mean Proportion	Range
Person	.034	.034	.060
Item	.047	.047	.093
Judge	.006	.006	.013
Occasion	.003	.003	.005
Person X Item	.180	.180	.301
Person X Judge	.011	.011	.012
Person X Occasion	.000	.000	.000
Item X Judge	.000	.000	.000
Item X Occasion	.000	.000	.000
Judge X Occasion	.006	.006	.013
Person X Item X Judge	.006	.006	.013
Person X Item X Occasion	.374	.374	.468
Person X Judge X Occasion	.028	.028	.046
Item X Judge X Occasion	.002	.002	.004
Person X Item X Judge X Occasion, Error	.305	.305	.526